Formal Request for Retraction of Published Article

Christine Stabell Benn, professor
Bandim Health Project, University of Southern Denmark
Studiestræde 6, 1455 Copenhagen K, Denmark
cbenn@health.sdu.dk; Tel: 45 2885 3964

December 10, 2025

Editors-in-Chief
Vaccine
Elsevier

Subject: Formal request for retraction of the article
Støvring H. et al. "What is actually the emerging evidence about non-specific vaccine effects in randomized trials from the Bandim Health Project"
Article ID: S0264410X25012344

Dear Editors-in-Chief,

We write to formally request the retraction of the above-referenced commentary by Henrik Støvring (HS) et al., based on serious scientific, methodological, and ethical deficiencies that fundamentally undermine its conclusions and its appropriateness for the scientific record.

This request is based on a detailed methodological review that highlights extensive errors, misrepresentation, misinterpretation, violations of fundamental statistical principles and misuse of statistical methods. This renders the commentary scientifically unreliable and potentially damaging to the research record.

The commentary not only contains material errors but presents them as the basis for public allegations of misconduct. HS et al. advance extremely serious allegations against researchers from the Bandim Health Project (BHP), including insinuations of "questionable research practice", "selection bias", and "p-hacking". Such claims against specific people, Christine Stabell Benn (CSB), Peter Aaby (PA) and a specific research group, demand the highest standards of accuracy, transparency, and evidential support. These standards are demonstrably not met in the published commentary.

Because the scientific subject is vaccines and their potential non-specific effects, inaccurate and misleading analyses can have direct consequences for public health understanding. It is therefore crucial that such analyses meet the highest standards of scientific rigor and integrity.

We believe the extent and seriousness of the critical errors is sufficient to justify a "Retraction" of the paper.

A detailed review of the deficiencies is set out below.

A. Study Methodology and Scientific Validity

A.1. Faulty application of Z-curve methodology

HS et al. applied Z-curve to more than 1,400 statistical tests drawn from 26 publications representing only 13 unique randomized controlled trials (RCTs). This approach violates core methodological principles required for valid Z-curve inference and systematically biases the estimates.

The key problems are as follows:

A.1.1. The Z-curve 2.0 methodology requires *many independent studies*, not many tests from a few studies

The Z-curve is designed to estimate evidential value across a set of *studies*, not to aggregate thousands of dependent tests from a small number of datasets[1] [2]. When many tests come from the same underlying data, the tests are not independent.

- The appropriate correction is a clustered/hierarchical bootstrap, where the *study* (here RCT) is the unit of resampling. However, this requires a large number of studies (clusters, ideally at least 100); with only 13 unique RCTs, the estimates are *inherently unstable*.

Conclusion: Using many hundreds of dependent tests from 13 RCTs violates the basic requirement of sufficient independent clusters for the Z-curve analysis to produce reliable estimates.

A.1.2. HS et al. heavily overweight single RCTs by including hundreds of nearly redundant tests within the same RCTs

Several papers of the RCTs included report dozens or even hundreds of tests on:

- the *same outcome*
- from the *same participants*
- with trivial variations (unadjusted and adjusted, intention-to-treat (ITT) and per protocol (PP), multiple time windows, overlapping subgroups, multiple interaction terms)

For example, one paper (PaperID=18) contributes 168 tests for the *same secondary outcome* (hospitalizations for infections). This is because it presents both crude and adjusted estimates from both ITT and PP analyses. These are not distinct hypotheses; they are repeated analyses of the same underlying effect, to test the robustness across various model specifications. Including these as 168 independent test statistics:

- artificially inflates the apparent sample size
- grossly overrepresents a single RCT's influence
- distorts the z-score distribution that the Z-curve fits

Conclusion: This usage contradicts best practice: only substantively distinct hypotheses should be included, with one test per primary outcome. Repeated model variants should not be counted as separate evidence.

A.1.3. The selection of tests is not aligned with any coherent research question

Z-curve requires that test selection reflects a well-defined research question. According to methodological standards:

- If the question is evidential strength, one typically examines the primary outcomes (those that were powered and prespecified)

[1] Bartoš F, Schimmack U. Z-curve 2.0: Estimating Replication Rates and Discovery Rates. Meta-Psychol 2022;6.
[2] Schimmack U, Bartoš F. Estimating the false discovery risk of (randomized) clinical trials in medical journals based on published p-values. PLoS One. 2023 Aug 30;18(8):e0290084

- If the question is research practices or bias, then secondary outcomes may be included, but must be analyzed with clustered inference and clear justification

HS et al. mix:

- primary outcomes,
- numerous secondary outcomes,
- exploratory subgroup analyses,
- interaction terms
- alternative model specifications

This violates the principle that only relevant and conceptually distinct tests should be included.

Conclusion: The data aggregated by HS et al is methodologically incoherent and not appropriate for a Z-curve analysis intended to evaluate either evidential value or research practices.

Conclusion A.1. HS et al. have applied the Z-curve methodology in a way that conflicts with established methodological principles for test selection, study-level independence, and appropriate clustering. Their dataset:

- contains far too few independent studies to support stable estimation,
- massively overweighs individual RCTs by including hundreds of redundant tests,
- mixes primary, secondary, exploratory, and model-specification variants without a coherent rationale, which introduces violations of selection assumptions that Z-curve cannot correct.

Therefore, the resulting ODR, EDR, and ERR estimates cannot be interpreted as valid evidence about evidential value or research practices in the underlying trials.

A.2. Incorrect representation and misinterpretation

HS et al mischaracterize the BHP's publications as making definitive claims of causal non-specific effects (NSEs). This assertion is demonstrably incorrect, as documented below and further substantiated in the appendix table. The BHP papers clearly distinguish between primary, secondary, and exploratory analyses and apply cautious language where appropriate. Supplementary analyses and subgroup findings are presented as tentative and exploratory. **The selective portrayal of these studies constructs a narrative of misconduct that is not supported by the source material and undermines norms of fair scientific critique.**

A.2.1. Incorrect representation and misinterpretation of primary outcomes

In their Table 1, HS et al. present the results on NSEs in the included papers. In criterion "2a" it is stated that only 1 of 12 papers reporting a primary outcome is significant [There are actually 13 studies reporting a primary outcome, since PaperID=1 also reports a primary outcome]. The "2a" count is factually incorrect since 3 other studies had significant primary findings based on their prespecified per protocol analysis (as approved by the ethical committee):

- PaperID: 4: Had significant primary outcome with a one-sided statistical test as was prespecified in the study protocol, p=0.04. [However, one reviewer obliged us to only report the 2-sided CI]

- PaperID; 28: Had a significant primary outcome in the per protocol analysis, HR=0.70 (0.52-0.94), p=0.02. [The protocol did not specify whether the per-protocol (PP) or the intention-to-treat (ITT) analysis was primary; since the new aspect of the RCT was two doses of measles vaccine, we emphasized the PP-analysis].

- PaperID: 36: HS et al. overlook that there are 2 primary outcomes. One is significant (also corrected for multiplicity among the primary outcomes), HR=0.50 (95% CI: 0.32-0.80), p=0.003 – see example below:

Effect of early two-dose measles vaccination on childhood mortality and modification by maternal measles antibody in Guinea-Bissau, West Africa: A single-centre open-label randomised controlled trial

Sebastian Nielsen,[a,b] Ane B Fisker,[a,b] Isaquel da Silva,[a] Stine Byberg,[a] Sofie Biering-Sørensen,[a] Carlitos Balé,[a] Amarildo Barbosa,[a]

Background Early 2-dose measles vaccine (MV) at 4 and 9 months of age vs. the WHO strategy of MV at 9 months of age reduced all-cause child mortality in a previous trial. We aimed to test two hypotheses: 1) a 2-dose strategy reduces child mortality between 4 and 60 months of age by 30%; 2) receiving early MV at 4 months in the presence versus absence of maternal measles antibodies (MatAb) reduces child mortality by 35%.

OPV before and after enrolment (p for interaction=0.027) [deaths/children: $n_{2\text{-dose}}$=27/1,602; $n_{1\text{-dose}}$=3/837]. In the 2-dose group receiving early MV at 4 months, mortality was 50% (20–68%) lower for those vaccinated in the presence of MatAb vs. the absence of MatAb [deaths/children: n_{MatAb}=51/3,132; $n_{noMatAb}$=31/1,028].

Conclusion: HS et al. make numerous errors in the presentation of our findings on the primary outcome(s) of the RCTs. These misrepresentations materially mislead readers and invalidate the central narrative of the commentary. See also table in appendix.

A.2.2. Incorrect representation and misinterpretation of secondary outcomes

In Table 1, criterion "3" HS et al. claim that among the 25 papers they assess as having no significant primary outcome, only **one** secondary outcome remained significant after applying the Holm-Bonferroni (HB) correction for multiplicity of tests. This is, however, not true, and HS et al. have not interpreted their own results correctly. There are multiple internal contradictions in supplementary table 2: The column "Holm-B significant p-value?" has 8 studies with a "Yes"

(PaperIDs: 4, 6, 10, 12, 13, 22, 32, 36). However, the column “HB results support secondary findings of NSE”, says “No” for these studies.

Example: PaperID 4: HS et al find 5 significant tests siginifcant after HB correction

Paper ID	Paper	origin	Publ used MC?	Estimate	LCL	UCL	Reported p-value	HB_pval	Holm-B significant p-value?
4	Biering-Sørensen, S., Aaby, P.,								No
4	Biering-Sørensen, S., Aaby, P.,	Table 2							No
4	Biering-Sørensen, S., Aaby, P.,	Table 3			0.57	0.35	0.93		0.02 Yes
4	Biering-Sørensen, S., Aaby, P.,	Table 3			0.37	0.17	0.84		0.03 Yes
4	Biering-Sørensen, S., Aaby, P.,	Figure 3			0.55	0.34	0.89		0.13 No
4	Biering-Sørensen, S., Aaby, P.,	Figure 3						0.03	0.11 No
4	Biering-Sørensen, S., Aaby, P.,	Figure 3						0.03	0.10 No
4	Biering-Sørensen, S., Aaby, P.,	Figure 3						0.00	0.01 Yes
4	Biering-Sørensen, S., Aaby, P.,	Figure 3						0.00	0.01 Yes
4	Biering-Sørensen, S., Aaby, P.,	Figure 3						0.05	0.05 Yes
4	Biering-Sørensen, S., Aaby, P.,	Figure 3						0.04	0.07 No

Yet they still conclude that the estimates are not significant after HB correction in the same table:

Paper reports primary results for RCT	*Results support primary finding of NSE (only relevant if N=Yes)*	*HB Results support primary finding of NSE (only relevant if N=No)*	*HB results support secondary findings of NSE*	*Authors primary conclusion is finding of a NSE*	*Authors secondary conclusion is finding of a NSE*	*Year of publication*	*Has trial ID*	*Analysis comparing randomization arms*
Yes	No		No	Yes	Yes	2017	Yes	Yes

Thus, according to HS et al.’s own results, 9 studies (including the one study acknowledged by HS et al.) support findings of NSEs based on the HB-corrected secondary outcomes.

Under criterion “5” HS et al. state that claims of NSEs are postulated by BHP in 23/25 papers based on secondary outcomes. HS conclude that NSEs were only shown for one study after HB correction. This is incorrect. Among the 22 remaining studies:

- 3/22 studies have a significant primary outcome and should therefore not be considered in this analysis (PaperID: 4, 28 and 36)
- 7/22 studies had a significant secondary NSE outcome after HB-correction (PaperID: 6, 10, 12, 13, 22, 32, 35)
- 12/22 papers did not have a significant secondary outcome after HB-correction but did also not claim proof of NSEs (PaperID: 1, 8, 9, 11, 16, 17, 18, 30, 31, 33, 39 and 40).
- Formulations such as the following:

- "… it may have…",
- "In subgroup analyses we found BCG re-vaccination might increase…"
- "... BCG tended to…"

do **not** claim causality and support of NSEs but rather indicates a potential association which needs to be investigated further in future research. See also the appendix table.

Example (PaperID=18):

BCG Vaccination at Birth and Rate of Hospitalization for Infection Until 15 Months of Age in Danish Children: A Randomized Clinical Multicenter Trial

Lone Graff Stensballe,[1] Henrik Ravn,[3] Nina Marie Birk,[4] Jesper Kjærgaard,[5] Thomas Nørrelykke Nissen,[4] Gitte Thybo Pihl,[6] Lisbeth Marianne Thøstesen,[6]

Conclusion. BCG vaccination did not affect the rate of hospitalization for infection up to the age of 15 months in Danish children. In future studies, the role of maternal BCG-vaccination, premature birth, and cesarean delivery needs further exploration.

Conclusion, A.2.: HS et al. made numerous errors in the presentation and interpretation of their own data and our data and analyses. For 8 papers they find significant evidence of NSEs after HB-correction on secondary outcomes but reach the opposite conclusion without any explanation, showing an extraordinary level of internal contradiction in their commentary. HS et al.'s claim furthermore misrepresents the BHP papers, by stating that in 23/25 of BHP studies, causal evidence for NSEs is claimed based on secondary outcomes without supporting statistical evidence for 22 of the studies. As documented above and in the appendix table, this is factually incorrect. These misrepresentations materially mislead readers and invalidate the central narrative of the commentary.

A.3. Erroneous extraction and inflation of number of tests in HB-correction

In direct violation of the principles for HB-correction (which controls family-wise error only if tests are independent (and relevant) for multiplicity correction), the authors correct for tests that do not form part of a hypothesis[3,4]:

Some examples are presented here (PaperID=6 and 32, respectively):

Double counted estimates and p-values

			All (n = 467)	
		Obs in Range	GMR (95% CI)	*P* Value
Medium	IL-1β	85%	1.33 (.97–1.83)	.08
	IL-6	87%	1.27 (.90–1.80)	.17
	TNF-α	96%	**1.30 (1.05–1.60)**	**.01**
	IL-5	60%	0.98 (.77–1.25)	.86
	IL-10	76%	1.13 (.85–1.52)	.40
	IL-17	62%	1.00 (.73–1.38)	.98
	IFN-γ	82%	**1.40 (1.04–1.88)**	**.03**

Counted both estimate and reference estimate of 1.00 as test

Table 2 Mortality and HRs for death among children eligible for enrolment in the MVEPI trial. Overall and eligibility assessment

	N	Deaths/person years (PYRS)	Mortality rate (per 1000 PYRS)	HR 95% CI*
Children 12–35 months at eligibility assessment				
Restrictive MV policy	1373	44/3698	11.9	1.00 (ref)
MV for all	1405	45/3723	12.1	0.95 (0.64 to 1.43)
All Children				
Restrictive MV policy	2339	81/6775	12.0	1.00 (ref)
MV for all	2428	92/6983	13.2	1.06 (0.78 to 1.44)

A full list of types of incorrectly extracted tests with examples (these errors are presumably also erroneously included in the Z-curve and other analyses in the commentary):

- Randomization checks and baseline characteristic comparisons (e.g., PaperID=4 includes 16 p-values derived solely from baseline characteristic comparisons (Table 1), all of which were incorrectly treated as independent hypothesis tests)
- Quality control and robustness analyses (e.g., PaperID=35 has 18 instances of excluding measles admissions to assess how much of the observed effect of measles vaccine was unrelated to measles (Tables 1 and 3))
- Exploratory analyses clearly labelled as such (e.g., PaperID=32 has a full table devoted to explorative analyses, these are still counted towards the total number of tests)
- Duplicated tests (both estimate and corresponding p-value)(e.g., PaperID=6 has 42 of these duplicate tests (Table 2))
- Multiple representations of the same data (e.g. unadjusted/adjusted, ITT/PP effect estimates and corresponding p-values)(e.g., PaperID=18 presents both crude and adjusted p-values for *per protocol* and ITT analyses; all are counted as independent tests)
- Reference group estimates (e.g. values fixed at 1.00) miscounted as separate tests (e.g., PaperID=32 has 14 of these estimates in Tables 2-4)

Conclusion A.3: The application of HB and other multiplicity corrections is fundamentally flawed with the present data. This inclusion of numerous dependent estimates and p-values falsely inflates the test count and leads to over-correction, artificially diluting statistically significant findings and distorting the scientific interpretation[5].

[3] Sarkar SK, Fu Y, Guo W. Improving Holm's procedure using pairwise dependencies. *Biometrika* 2016; 103(1): 237-43.
[4] Stevens JR, Al Masud A, Suyundikov A. A comparison of multiple testing adjustment methods with block-correlation positively-dependent tests. *PLOS ONE* 2017; **12**(4): e0176124
[5] Dmitrienko A; Chapman & Hall/CRC; 2010

A.4. Unclear and flawed selection of papers

HS et al's commentary is unusual because it does not aim to investigate a research question, e.g., what is the evidence for NSEs of vaccines? It zooms in on a research group, and a subset of studies (RCTs) from that group. Thus, all the evidence for NSEs from other studies and other groups is ignored. It is hard to understand the purpose and see the relevance for the broader research community. Even within this narrow scope, the commentary has not included all relevant papers. While HS et al. do not claim to have included all relevant papers from the BHP, they "are confident that our research has retrieved the most important ones". However, several of BHP's most important RCTs documenting NSEs were either not identified in the first place or excluded. For example, the RCTs of the high-titer measles vaccine (HTMV) that led WHO to withdraw the vaccine (PaperID=23; 24; 25; 26) were excluded because they were conducted prior to the time when RCTs were pre-registered. The HTMV study is the clearest case of NSEs of a vaccine; HTMV was protective against measles infection but associated with increased female mortality. This selective exclusion disproportionately removes studies demonstrating strong NSEs, thereby biasing the analysis toward underestimating or negating true effects. Using a criterion of only including pre-registered trial would dismiss most of the basis for modern medicine.

Conclusions on Study Methodology and Scientific Validity based on A.1.-A.4.

We have documented the following critical errors in the commentary by HS et al.

- **Invalid statistical foundation:** Use of Z-curve analysis on highly dependent tests across and within papers, invalidating the entire analysis and its conclusions.
- **Systematic misinterpretation:** HS et al. misrepresent and misinterpret the analysis of the primary and secondary outcomes. They fail to report their own results from the HB-correction truthfully, since many papers have significant secondary outcomes even after HB-correction, but are reported by HS et al. as having no significant outcomes after HB-correction.
- **Misrepresentation of BHP claims:** HS et al. do not truthfully present the papers by BHP. E.g., HS et al. report that we claim findings of NSEs based on secondary outcomes that do not stay significant after HB-correction, but this is wrong, either because there is indeed a significant secondary outcome even after HB correction, or because no claim of proof of NSEs were made by the BHP authors.
- **Inflated test counts:** HS et al. make hundreds of mistakes when extracting tests for the HB-correction, this inflates the number of tests and dilutes the chance of finding significant findings.
- **Arbitrary study selection:** HS et al.'s paper is based on an arbitrary and flawed selection of papers.

All these critical errors undermine not only the results but also the conclusions made by HS et al.

B. Defamatory Claims

The commentary contains repeated statements implying misconduct, including allegations of outcome switching, cherry-picking, reinterpretation of trials, and "p-hacking". These serious accusations are made without adequate documentation or evidential support and are directed at named individuals. These statements meet commonly accepted definitions of reputational harm by publicly imputing dishonesty and unethical conduct.

Requested Action

In accordance with Elsevier's retraction policy and COPE guidelines, we respectfully request that *Vaccine* retracts the commentary and issue a formal editorial notice clearly stating the reasons for retraction.

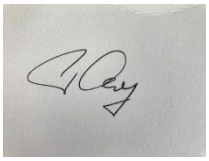
The shortcomings documented above meet COPE retraction criteria for unreliable findings, inappropriate analytic methodology, and unsubstantiated defamatory assertions.

We expect the editorial board to act promptly in accordance with Elsevier and COPE guidelines and to inform us of the timeline for its decision.

Yours sincerely,

Christine Stabell Benn

Professor, MD, PhD, DMSc



Peter Aaby

Professor, MSc, DMSc

Sebastian Nielsen,

Senior Statistician, MSc, PhD

Appendix: Overview of the findings and claimed findings by HS et al and BHP, respectively, for the 26 papers presented in HS et al's paper.

Table 1A. 7 papers in which HS et al state that the BHP authors have analyzed a primary/main outcome and claimed proof of NSEs.

Paper ID	BHP claims proof of NSEs based on primary or main outcomes?		Significant primary or main outcomes	HS et al. correct that BHP claims proof of NSEs without evidence
	According to HS et al.	According to BHP		
4	Yes	Yes	Yes	**No**
6	Yes	N/A #1	N/A #1	**No**
20	Yes	Yes	Yes	**N/A #2**
30	Yes	No	No	**No**
31	Yes	No	No	**No**
35	Yes	No	No	**No**
40	Yes	No	No	**No**

#1 This paper presents exploratory analyses with no clear definition of a primary or main outcome.
#2 Both parties agree that there was a statistically significant primary outcome.

Table 1B. 23 papers in which HS et al state that the BHP authors have analyzed secondary outcomes and claimed proof of NSEs.

Paper ID	BHP claims proof of NSEs based on secondary outcomes?		Significant secondary outcome after HB correction #3	HS et al. correct that BHP claims proof of NSEs without evidence after HB correction #3
	According to HS et al.	According to BHP		
1	Yes	No	No	**No**
4	Yes	Yes	Yes	**No #4**
6	Yes	Yes	Yes	**No**
8	Yes	No	No	**No**
9	Yes	No	No	**No**
10	Yes	Yes	Yes	**No**
11	Yes	No	No	**No**
12	Yes	Yes	Yes	**No**
13	Yes	Yes	Yes	**No**
16	Yes	No	No	**No**
17	Yes	No	No	**No**
18	Yes	No	No	**No**
22	Yes	Yes	Yes	**No**
28	Yes	Yes	No	**No #4**
30	Yes	No	No	**No**
31	Yes	No	No	**No**
32	Yes	Yes	Yes	**No**
33	Yes	No	No	**No**
35	Yes	No	No	**No**
36	Yes	Yes	Yes	**No #4**
37	Yes	Yes	Yes	**N/A #5**
39	Yes	No	No	**No**
40	Yes	No	No	**No**

#3 Bonferroni-Holm correction for multiple tests; #4 The primary outcome was significant; #5 Both parties agree there was significant effect after HB correction.